Stanford CS 224n Assignment 4

Yunxiang Zhang

March 31, 2021

1 Neural Machine Translation with RNNs

(g) **Answer:** In the `step()` function, the padded locations in e_t are set to $-\inf$. After applying `softmax()` function to e_t, the attention scores for 'pad' tokens are zero ($e^{-\inf} = 0$). As a result, the hidden states of 'pad' tokens do not contribute to the calculation of the attention output a_t. The masking operation is necessary because the 'pad' tokens contain no useful information for translation and we should not pay attention to them when calculating the attention output.

(h) **Answer:** The corpus BLEU score is 12.60.

(i) i. **Answer:** Advantage: the dot product attention introduces no extra parameters and is more efficient to compute. Disadvantage: it requires that s_t and h_i should have same dimensions, which is less versatile than multiplicative attention.

ii. **Answer:** Advantage: additive attention performs better for larger dimensions. Disadvantage: multiplicative attention is faster and more space-efficient in practice as it can be implemented more efficiently using matrix multiplication. Reference: Sebatian Ruder's blog post

2 Analyzing NMT Systems

(a) **Answer:** According to the Wikipedia page of polysynthetic language, they are languages in which words are composed of many morphemes (word parts that have independent meaning but may or may not be able to stand alone). To better encode the information of different components of Cherokee words, the subword-level embeddings are more suitable than word-level. Besides, subword-level model can handle words that does not appear in corpus, while word-level cannot.

(b) **Answer:** Different combinations of characters and subwords produce different words. Therefore, the total number of characters and subwords are less than the whole words.

(c) **Answer:** Massive multilingual models are effective at generalization, and capable of learning shared representations for linguistically similar languages without the need for external constraints. The learning signal from high-resource language, e.g. French to English, can be be transferred to benefit the translation quality of low-resource languages, e.g. Cherokee to English.

(d) i. **Answer:** Reason: the word *daisies* is a low frequency word and does not appear in the training corpus. The NMT Model cannot learn its meaning. Fix: Enlarge the training data to include less frequent words like *daisies*.

ii. **Answer:** Reason: the wrong pronoun "it" might be caused by insufficient representative capacity of the final softmax output with vocabulary projection. Fix: add more hidden layers to the vocabulary projection, which has only one layer currently.

iii. **Answer:** Reason: "Littlefish" refers to a special name here, but the NMT only captures its semantic meaning but does not remember the special form of the word. This error is likely caused by the low memory capacity of the model. Fix: add more hidden layers to encoder and decoder or tweak the size of them.

(e) i. **Answer:** Predicted translation: *"And the devil said unto him, If thou art the Son of God, command this stone that it become bread."* Target translation: *"And the devil said unto him, If thou art the Son of God, because the stone of this stone."* The training file does contain that string verbatim, e.g. *"but these are written, that ye may believe that Jesus is the Christ, the Son of God; and that believing ye may have life in his name."* It shows that the NMT system learns to "memorize" special phrases/combinations of words.

ii. **Answer:** Predicted translation: *"For this is the love of God, that we keep his commandments: and his commandments are not grievous."* Target translation: *"For this is the love of God, that he may send me his commandments; and his commandments is not."* This divergence shows that the model has a limited view of decoding due to the fixed beam search size. Although it successfully decodes *"the love of God"*, which appears verbatim in the training corpus, it heads in a wrong direction afterwards.

(f) i. **Answer:**

BLEU score for c_1

$p_1 = (0+1+1+1+0)/5 = 0.6, p_2 = (0+1+1+0)/4 = 0.5, len(c) = 5, len(r) = 4, BP = 1$

$BLEU_{c_1} = 1 \times \exp(0.5 \times \log 0.6 + 0.5 \times \log 0.5) = 0.55$

BLEU score for c_2

$p_1 = (1+1+0+1+1)/5 = 0.8, p_2 = (1+0+0+1)/4 = 0.5, len(c) = 5, len(r) = 4, BP = 1$

$BLEU_{c_2} = 1 \times \exp(0.5 \times \log 0.8 + 0.5 \times \log 0.5) = 0.63$

According to BLEU scores, c_2 is a better translation and I agree with the score.

ii. **Answer:**

BLEU score for c_1

$p_1 = (0+1+1+1+0)/5 = 0.6, p_2 = (0+1+1+0)/4 = 0.5, len(c) = 5, len(r) = 6, BP = \exp(1-6/5) = 0.8187$

$BLEU_{c_1} = 0.8187 \times \exp(0.5 \times \log 0.6 + 0.5 \times \log 0.5) = 0.45$

BLEU score for c_2

$p_1 = (1+1+0+0+0)/5 = 0.4, p_2 = (1+0+0+0)/4 = 0.25, len(c) = 5, len(r) = 6, BP = \exp(1-6/5) = 0.8187$

$BLEU_{c_2} = 0.8187 \times \exp(0.5 \times \log 0.4 + 0.5 \times \log 0.25) = 0.26$

c_1 now receives the higher BLEU score and I **do not** agree with the score.

iii. **Answer:** There are multiple ways of translating a sentence, sometimes with little overlap. With only a single reference, the BLEU score for a candidate translation is biased and cannot represent the quality of translation comprehensively.

iv. **Answer:**

advantages:

- BLEU score can be computed automatically, which is faster than human evaluation
- BLEU score is objective, while human evaluation can be influenced by personal understandings of the source and target languages.

disadvantages:

- If the number of reference translations is limited, BLEU score cannot reflect the quality of translated sentence properly, because there can be multiple ways of translation for a single sentence.
- BLEU score does not consider whether the meaning and grammar are correct for translated sentences.